Journal of Multivariate Analysis 73, 18–40 (2000)
doi:10.1006/jmva.1999.1861, available online at http://www.idealibrary.com on IDEAL®

Estimation of Densities and Derivatives of Densities with Directional Data[1]

Jussi Klemelä

Rolf Nevanlinna Institute, University of Helsinki,
P.O. Box 4, SF-00014 Helsingin Yliopisto, Finland

Received March 20, 1997

Estimating the density function of a random vector taking values on the d-dimensional unit sphere is considered. Also the estimation of the Laplacian of the density

Watson and Cabrera (1987). It is also proved that asymptotically the plug-in method is as good as using the asymptotically optimal deterministic smoothing parameter sequence. © 2000 Academic Press

AMS 1991 subject classifications: 62G07, 62H11.

Key words and phrases: kernel estimator, Laplace operator, L_p error, nonparametric density estimation, plug-in method, spherical data.

1. INTRODUCTION

When statistical data consists of directions, it can be represented as points of the unit sphere in the Euclidean space. The notation $S_d = \{x \in \mathbf{R}^{d+1} : \|x\| = 1\}$, $d \geqslant 1$, is used but only the case $d \geqslant 2$ will be considered. From a practical point of view, the case $d = 2$ is the most important, but see Diaconis (1988, p. 100) for the case where $d \geqslant 3$. Good introductions to the statistics of spherical data are given by Mardia (1972), Watson (1983) and Fisher, Lewis and Embleton (1987).

It is assumed that $X_1, ..., X_n$ are independent identically distributed observations with values on S_d, and their distribution is absolutely continuous with respect to the Lebesgue measure of the sphere, which will be denoted by $\mu = \mu_d$. It will be denoted $\omega_d \stackrel{\text{def}}{=} \mu(S_d) = 2\pi^{(d+1)/2}/\Gamma((d+1)/2)$. We will consider the estimation of the density of the observations with kernel estimator. The kernel estimator of a density will be defined in such a way

[1] This work was supported in part by the Academy of Finland under Grant no. 1015998 and by the University of Helsinki under a research grant.

0047-259X/00 $35.00

that the value of the estimator at $x \in S_d$ will depend only on the distance between x and the observations. The distance will be the Riemannian distance $\delta(x, y) = \arccos(x'y)$.

DEFINITION 1.1. Kernel density estimator with the kernel function $L: [0, \infty[\to \mathbf{R}$ and the smoothing parameter $\kappa > 0$ is

$$\hat{f}_n(x) = \hat{f}_n(x, \kappa, L) = \frac{c(\kappa)}{n} \sum_{i=1}^{n} L(\kappa \arccos(x'X_i)),$$

where $x \in S_d$ and

$$c(\kappa)^{-1} = \int_{S_d} L(\kappa \arccos(x'y))\, d\mu(y). \tag{1}$$

Reasonable choices for the kernel function are for example $L(t) = e^{-t^2} I_{[0, \infty[}(t)$, $L(t) = (1 - t^2) I_{[0, 1]}(t)$ and $L(t) = I_{[0, 1]}(t)$. The choice $L(t) = I_{[0, 1]}(t)$ leads to the so called naive estimator, $\hat{f}_n(x) = P_n(C_\kappa(x))/\mu(C_\kappa(x))$, where $P_n(C_\kappa(x)) = n^{-1} \sum_{i=1}^{n} I_{C_\kappa(x)}(X_i)$ and $C_\kappa(x) = \{ y \in S_d \mid x'y \leqslant \cos(\kappa^{-1}) \}$. The naive estimator was studied by Ruymgaart (1989), Hendriks, Janssen and Ruymgaart (1993).

In Watson (1970), Beran (1979), Hall, Watson and Cabreira (1987), Bai, Radhakrishna Rao and Zhao (1988) the form $n^{-1} c(\kappa) \sum_{i=1}^{n} L(\kappa^2(1 - x'X_i))$ has been considered, where $L: [0, \infty[\to \mathbf{R}$ and $c(\kappa)$ is the normalization constant. Note that $1 - x'X_i = \|x - X_i\|^2/2$. In Hall, Watson and Cabrera (1987), also the form $K_\kappa(x) = c(\kappa)\, J(\kappa^2 x'\eta)$ has been considered, where $J: \mathbf{R} \to \mathbf{R}$. It was shown that asymptotically the latter form is a special case of the first form.

Watson (1970) contained the definition of estimator and Beran (1979) was interested from using kernel estimator in constructing robust estimators for parameters. An extensive study of statistical properties of the kernel estimator was given in Hall, Watson and Cabrera (1987), where rates of convergence were established for pointwise, L_2, and Kullback–Leibler loss, for 2-smooth densities. Furthermore, they considered cross-validation as a method for choosing the smoothing parameter. Bai, Radhakrishna Rao and Zhao (1988) gave conditions of pointwise strong consistency, uniform strong consistency, and L_1-norm consistency of the kernel estimator.

In this article the results of Hall, Watson and Cabrera (1987) are generalized to cover estimation of s-smooth densities, $s \geqslant 2$. The pointwise, L_2, and L_1 losses are covered. It is shown that the asymptotics of the bias

does not involve the Laplacian of the density when $s \geqslant 4$, but another type of derivative. The plug-in method is used to choose smoothing parameter empirically. It is proved that the resulting L_2 risk is asymptotically the same as when using the asymptotically optimal deterministic smoothing parameter.

We consider also estimation of the Laplacians and other type of derivatives. In the Euclidean case, derivatives of the kernel estimator are also kernel estimators. In the spherical case, however, the Laplacian of the kernel estimator is not a kernel estimator. Thus we have two natural estimators for the Laplacian of the density. The other is the kernel estimator and the other is the Laplacian of the kernel estimator. Rates of convergence are given for the pointwise risk.

In Section 2 we discuss how the kernel estimators arise as a special case of delta sequence estimators. Then we define an estimator for an iterated Laplacian of the density as an iterated Laplacian of the kernel estimator. Then definition of other type of derivative is given and an estimator for this derivative is defined as a linear combination of kernel estimators. In Section 3 the bias of the estimators is studied. The expectation of the kernel estimator is a convolution and thus Section 3 is concerned with the study of approximation by convolution. In Section 4 the results about the asymptotics of the pointwise, L_2, and L_1 risk are given. In Section 5 in a certain sense asymptotically optimal choice for the smoothing parameter is given.

The proofs which are not given in this article can be found from Klemelä (1997).

Let us give the definitions of two parameterizations of S_d. Let $\eta \in S_d$ and $T_\eta = \{\xi \in S_d \mid \xi \perp \eta\}$. Let $\phi_\eta \colon S_d \backslash \{\eta, -\eta\} \to T_\eta \times \,]0, \pi[$ be a parameterization of S_d defined by

$$\phi_\eta{}^{-1}(\xi, \theta) = \eta \cos \theta + \xi \sin \theta. \tag{2}$$

Note that $\phi_\eta{}^{-1}(\xi, \theta)$ is well defined for all $\theta \in \mathbf{R}$, although there does not exist any function for which $\phi_\eta{}^{-1} \colon T_\eta \times \mathbf{R} \to S_d$ would be an inverse function. The second possibility is to denote by ϕ_η a mapping $S_d \backslash \{\eta, -\eta\} \to T_\eta \times \,]-1, 1[$, defined by

$$\phi_0^{-1}(\xi, t) = \eta t + \xi(1 - t^2)^{1/2}. \tag{3}$$

To the first parameterization corresponds the integration formula

$$\int_{S_d} f(x)\, d\mu(x) = \int_0^\pi d\theta \sin^{d-1} \theta \int_{T_\eta} f(\phi_\eta^{-1}(\xi, \theta))\, d\mu_{d-1}(\xi).$$

2. DELTA SEQUENCE ESTIMATORS FOR DENSITIES AND DERIVATIVES

If $X_1, ..., X_n$ are independent identically distributed random variables with values in $\mathbf{R}^d$, $d \geqslant 1$, an estimator for their common density is

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_\kappa(x - X_i),$$

where $x \in \mathbf{R}^d$ and $K_\kappa: \mathbf{R}^d \to \mathbf{R}$ is a function which depends on a positive parameter κ and whose mass concentrates more and more in the vicinity of the origin as $\kappa \to \infty$. By this we mean that $\lim_{\kappa \to \infty} \int_{\|y\| > \delta} |K_\kappa(y)|\, dy = 0$ for every $\delta > 0$. If also $\lim_{\kappa \to \infty} \int_{\mathbf{R}^d} K_\kappa = 1$, then the term approximate identity or delta sequence has been used for such sequences of functions. This estimator, which could be called a delta sequence estimator, was studied by Watson and Leadbetter (1963). The delta sequence estimator is called kernel estimator if $K_\kappa(x) = \kappa^d K(\kappa x)$, where $K: \mathbf{R}^d \to \mathbf{R}$ is a kernel function (typically a density function) and $\kappa > 0$ is a smoothing parameter.

The delta sequence estimator can be defined also in the case of spherical data. Let $\eta \in S_d$ be fixed and assume that we have constructed such functions $K_\kappa: S_d \to \mathbf{R}$ for $\kappa > 0$ that the mass of K_κ concentrates more and more in the vicinity of η, when $\kappa \to \infty$. For $x, y \in S_d$, let $R_{x,y}: S_d \to S_d$ be any such rotation that $R_{x,y}(x) = y$ (rotation is a restriction to S_d of a linear map whose matrix is orthogonal with determinant one). If $X_1, ..., X_n$ are i.i.d random variables with values on S_d, the delta sequence estimator of their common density can be defined as

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_\kappa(R_{X_i,\eta}(x)), \tag{4}$$

where $x \in S_d$. This definition replaces translation in Euclidean space by rotation on the sphere. It does not matter which version of the rotations we have chosen. The function K_κ is "centered" on each observation in turn and the average over the observations is taken. An essentially equivalent way of defining the delta sequence estimator in the case of Euclidean data is

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_\kappa(X_i - x).$$

In the case of spherical data we could have defined

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_\kappa(R_{x,\eta}(X_i)). \tag{5}$$

In the case of Euclidean data, sometimes a restriction is made to consider only delta sequences of the form $K_\kappa(x) = \kappa^d L(\|\kappa x\|)$, where $x \in \mathbf{R}^d$ and $L: [0, \infty[\to \mathbf{R}$. These functions depend only on the distance between the argument and the origin. Similarly in the case of spherical data we get a large class of delta sequences when restricting ourselves to functions which depend only on the Riemann distance between the argument and the vector η. This study will be restricted in such a way to the case of $K_\kappa(x) = c(\kappa)\, L(\kappa \arccos(x'\eta))$. For this case it holds that $K_\kappa(R_{X_i, \eta}(x)) = c(\kappa)\, L(\kappa \arccos(x' X_i)) = K_\kappa(R_{x, \eta}(X_i))$ and thus in this case the definitions (4) and (5) can be simplified to the Definition 1.1.

We will consider not only estimation of a density but also estimation of the Laplacian of a density. Let us define the Laplace operator. Let $f: S_d \to \mathbf{R}$ and let $\eta^1, ..., \eta^{d+1} \in S_d$ be orthogonal. The Laplace operator is defined recursively by

$$\begin{aligned} \Delta f(x) &= \Delta_d f(x) \\ &= \left[(1-t^2) \frac{\partial^2}{\partial t^2} - td \frac{\partial}{\partial t} + \frac{1}{1-t^2} \Delta_{d-1} \right] f(\phi_{\eta^{d+1}}^{-1}(\xi, t))|_{(\xi, t) = \phi_{\eta^{d+1}}(x)} \end{aligned}$$

for $d \geqslant 2$ where $\phi_{\eta^{d+1}}^{-1}$ was defined in (3). Define $\Delta_1 f(x) = \partial^2/\partial\theta^2\, f(\phi_{\eta^1, \eta^2}^{-1}(\theta))|_{\theta = \phi_{\eta^1, \eta^2}(x)}$ where $\phi_{\eta^1, \eta^2}: S_1 \backslash \{\eta^2\} \to]0, 2\pi[$, $\eta^1, \eta^2 \in S_1$, $\eta^1 \perp \eta^2$, is defined by $\phi_{\eta^1, \eta^2}^{-1}(\theta) = \eta^2 \cos\theta + \eta^1 \sin\theta$. For $r \geqslant 4$ even, define $\Delta^{r/2} f = \Delta \Delta^{(r-2)/2} f$.

We will next construct an estimator for $\Delta^{r/2} f(x_0)$, $r \geqslant 0$ even. The estimator to be defined is the iterated Laplacian of the kernel estimator.

Let $L: [0, \infty[\to \mathbf{R}$ be such that $\int_0^\infty t^{d-1} |L(t)|\, dt < \infty$ and L has r derivatives. Define

$$\hat{g}_n(x) = \Delta^{r/2}(\hat{f}_n(x, \kappa, L)) = \frac{c(\kappa)}{n} \sum_{i=1}^{n} L^{(r, \kappa)}(x' X_i), \tag{6}$$

where $\hat{f}_n$ was defined in Definition 1.1, $L^{(r, \kappa)}: [-1, 1] \to \mathbf{R}$ is defined by $L^{(0, \kappa)}(t) = L(\kappa \arccos t)$, and

$$L^{(r, \kappa)}(t) = \left[(1-t^2) \frac{\partial^2}{\partial t^2} - dt \frac{\partial}{\partial t} \right] L^{(r-2, \kappa)}(t)$$

if $r \geqslant 2$. When $r = 0$, $\hat{g}_n$ is the usual kernel estimator of a density.

For example, $L^{(2, \kappa)}(t) = \kappa^2 L^{(2)}(\kappa \arccos t) + \kappa(d-1)\ \ L^{(1)}(\kappa \arccos t)\ (1-t^2)^{-1/2}\, t$ and $L^{(4, \kappa)}(t) = \kappa^4 L^{(4)}(\kappa \arccos t) + \kappa^3(d-1)\ \ L^{(3)}(\kappa \arccos t)$

$(1-t^2)^{-1/2}\,2t+\kappa^2(d-1)\,L^{(2)}(\kappa \arccos t)(1-t^2)^{-1}[(d-1)\,t^2-2]+\kappa(d-1)$ $L^{(1)}(\kappa \arccos t)(1-t^2)^{-3/2}\,t(3-d)$. Generally, for $r \geqslant 2$ even, it holds that

$$L^{(r,\kappa)}(t)=\sum_{i=1}^{r} \kappa^i L^{(i)}(\kappa \arccos t)(1-t^2)^{(r-i)/2}\,P_{r,i}(t), \tag{7}$$

where $P_{r,i}$ are polynomials, $P_{r,r}\equiv 1$, and when $d=1$, $P_{r,i}\equiv 0$ for $i=1, ..., r-1$. Thus, only when $d=1$ (or $r=0$), $\hat{g}_n$ has the form of a kernel estimator.

It will be seen that the higher order asymptotics of the kernel estimator, and more generally, the quality of the approximation of a function with a convolution depends not on the iterated Laplacian of the density but on the derivatives of other type. Let us define this other type of derivative.

When $g: \mathbf{R}^{d+1} \to \mathbf{R}$ and $x, \xi \in \mathbf{R}^{d+1}$, define the derivative of g at x in the direction of ξ to be $D_\xi g(x)=\lim_{h\to 0} h^{-1}[\,g(x+h\xi)-g(x)\,]$ and $D_\xi^s g = D_\xi D_\xi^{s-1} g$, for $s \geqslant 2$ integer. The derivative of $f: S_d \to \mathbf{R}$ of order s will be defined as "the average" of directional derivatives of order s, taken over all directions orthogonal to x.

DEFINITION 2.1. Let $f: S_d \to \mathbf{R}$ and define $\bar{f}: \mathbf{R}^{d+1} \to \mathbf{R}$ by $\bar{f}(x)=f(x/\|x\|)$. The derivative of f of order s is $D^s f: S_d \to \mathbf{R}$ defined by

$$D^s f(x)=\omega_{d-1}^{-1} \int_{T_x} D_\xi^s \bar{f}(x)\, d\mu_{d-1}(\xi),$$

where $d \geqslant 2$, $T_x=\{\xi \in S_d : \xi \perp x\}$, and $\omega_{d-1}=\mu_{d-1}(S_{d-1})$.

This concept was first defined by Hall, Watson and Cabrera (1987). Its relation to the Laplace operator is given in the next theorem.

THEOREM 2.2. *For $f: S_d \to \mathbf{R}$, $d \geqslant 2$,*

$$D^2 f = d^{-1}\,\Delta f,$$

if $\bar{f}$ and its partial derivatives are differentiable, where $\bar{f}(x)=f(x/\|x\|)$.

By considering a function of form $f: S_d \to \mathbf{R}$, $f(x)=g(x'\eta)$, where $\eta \in S_d$, $g: [-1, 1] \to \mathbf{R}$, it is seen that there exist no constant C for which $D^4 = C\Delta^2$.

Next an estimator for $\sum_{i=0}^{r/2} b_i D^{2i} f(x)$, $r \geqslant 0$ even, will be constructed, where $b_i \in \mathbf{R}$. Thus Δf will be estimated again, now with a different estimator. The estimator to be defined is a linear combination of kernel estimators.

Let $L_i: [0, \infty[\to \mathbf{R}$, $i = 0, ..., r/2$, be such that $\int_0^\infty t^{2i+d-1} |L_i(t)| \, dt < \infty$ and $\int_0^\infty t^{2i+d-1} L_i(t) \, dt \neq 0$. Let $a = (a_0, ..., a_{r/2}) \in \mathbf{R}^{r/2+1}$ and define

$$\hat{h}_{n,a}(x) = \frac{\kappa^d}{n} \sum_{i=1}^{n} M_{\kappa,a}(x'X_i) \tag{8}$$

where

$$M_{\kappa,a}(t) = \sum_{i=0}^{r/2} a_i C_i(L_i) \, \kappa^{2i} L_i(\kappa \arccos t), \qquad t \in [-1, 1],$$

$C_i(L_i) = [\omega_{d-1} \alpha_{2i}(L_i)]^{-1}$.

It will be seen in Lemma 3.6(ii) that $\hat{h}_{n,a}$ is an estimator for

$$\sum_{j=0}^{r/2} \frac{1}{(2j)!} D^{2j} f \sum_{i=j}^{r/2} \gamma_{i-j} a_i$$

where γ_i will be defined in (9). Thus, let a be such that

$$\frac{1}{(2j)!} \sum_{i=j}^{r/2} \gamma_{i-j} a_i = b_j, \qquad j = 0, ..., r/2.$$

For example, when $r = 0$, $b_0 = 1$, then $a_0 = 1$, and when $r = 2$, $(b_0, b_1) = (0, 1)$ then $(a_0, a_1) = (-2\gamma_1, 2) = (d/3, 2)$. When $r = 0$, estimator $\hat{h}_{n,a}$ is otherwise similar to the kernel estimator of a density but the normalization constant is different.

3. THE BIAS OF THE ESTIMATORS

For $L: [0, \infty[\to \mathbf{R}$, define $L_\kappa: [-1, 1] \to \mathbf{R}$ by $L_\kappa(t) = L(\kappa \arccos t)$, $\kappa > 0$. Define the convolution of $f: S_d \to \mathbf{R}$ and L_κ by

$$f * L_\kappa(x) = \int_{S_d} f(y) \, L_\kappa(x'y) \, d\mu(y)$$

$$= \int_{S_d} f(y) \, L(\kappa \arccos x'y) \, d\mu(y), \qquad x \in S_d.$$

For the kernel estimator $\hat{f}_n$,

$$\mathrm{E}(\hat{f}_n(x)) = c(\kappa) \, \mathrm{E}(L(\kappa \arccos(x'X_1)) = c(\kappa) \, f * L_\kappa(x).$$

Thus the study of convolutions is important for the study of the kernel estimator. Later, in Lemma 3.2, an expansion of convolution will be given.

The remainder term of this expansion is convenient to write in terms of an associated delta sequence, which will be defined next.

DEFINITION 3.1. Let $s \geqslant 0$ be integer. Let $L: [0, \infty[\to \mathbf{R}$ be such that $\int_0^\infty t^{s+d-1} |L(t)| \, dt < \infty$. The parameter s delta sequence associated with the delta sequence $c(\kappa) L(\kappa \arccos x'y)$ is $\tilde{L}_\kappa^{(s)}: [0, \pi] \to \mathbf{R}$, defined by

$$\tilde{L}_\kappa^{(s)}(\theta) = \omega_{d-1} \kappa^d \frac{\theta^s}{(s-1)!} \int_1^{\pi/\theta} (t-1)^{s-1} L(\kappa t \theta) \sin^{d-1}(t\theta) \, dt,$$

for $s \geqslant 1$ and

$$\tilde{L}_\kappa^{(0)}(\theta) = \omega_{d-1} \kappa^d L(\kappa\theta) \sin^{d-1}\theta.$$

The concept of an associated delta sequence is related to the concept of an associated kernel which is used in the Euclidean case. The associated kernel was defined for the one dimensional case in Bretagnolle and Huber (1979) and for the multivariate case in Holmström and Klemelä (1992). The following lemma gives an expansion of convolution.

LEMMA 3.2. *Let $x \in S_d$ and $s \geqslant 0$ even. Assume that for all $\xi \in T_x$, $\partial^s / \partial \theta^s \, f(\phi_x^{-1}(\xi, \theta))$ is continuous as a function of $\theta \in \mathbf{R}$. Let $\int_0^\infty t^{i+d-1} |L(t)| \, dt < \infty$ for $i = 0, s$. Then*

$$\kappa^d f * L_\kappa(x) = \sum_{i=0}^{s/2-1} \frac{1}{(2i)!} d_{2i}(\kappa, L) \, D^{2i} f(x) + \int_0^\pi \tilde{L}_\kappa^{(s)}(\theta) \, \tilde{D}^s f(x, \theta) \, d\theta,$$

where

$$d_{2i}(\kappa, L) = \omega_{d-1} \kappa^d \int_0^\pi \theta^{2i} L(\kappa\theta) \sin^{d-1}\theta \, d\theta$$

and $\tilde{D}^s f: S_d \times \mathbf{R} \to \mathbf{R}$ is defined by

$$\tilde{D}^s f(x, \theta) = \omega_{d-1}^{-1} \int_{T_x} D^s_{\phi_x^{-1}(\xi, \theta + \pi/2)} \bar{f}(\phi_x^{-1}(\xi, \theta)) \, d\mu_{d-1}(\xi),$$

where ϕ_x was defined in (2).

The function $\tilde{D}^s f(x, \theta)$ appearing in the previous lemma is again "average" over directions ξ orthogonal to x of certain directional derivatives. This time the directional derivatives are taken at points $\phi_x^{-1}(\xi, \theta)$ which are of distance determined by θ from x. Note that $\phi_x^{-1}(\xi, \theta + \pi/2) \perp \phi_x^{-1}(\xi, \theta)$.

In the expansion of Lemma 3.2 one would like that the kernel L could be chosen in such a way that $d_{2i}(\kappa, L) = 0$, for $i = 1, ..., s/2 - 1$. This is not possible but we can have $d_{2i}(\kappa, L)$ converge to zero arbitrarily fast, with a suitable choice of L. To prove this, we will give an expansion of the terms $d_{2i}(\kappa, L)$ in terms of the moments of L. Define

$$\alpha_s = \alpha_s(L) = \int_0^\infty t^{s+d-1} L(t)\, dt,$$

where $s \geqslant 0$. The next lemma gives an expansion of $d_m(\kappa, L)$ in terms of $\alpha_i(L)$.

LEMMA 3.3. *Let $m \geqslant 0$, $r \geqslant 0$ be integers and suppose that*

$$\int_0^\kappa t^{m+2i+d-1} L(t)\, dt = \alpha_{m+2i}(L) + o(\kappa^{2i-2r}),$$

for $i = 0, ..., r$. Then

$$d_m(\kappa, L) = \omega_{d-1} \sum_{i=0}^{r} \kappa^{-m-2i} \gamma_i \alpha_{m+2i}(L) + o(\kappa^{-m-2r}),$$

where

$$\gamma_i = \sum_{\alpha_1 + \cdots + \alpha_{d-1} = i} \frac{(-1)^{\alpha_1}}{(2\alpha_1 + 1)!} \cdots \frac{(-1)^{\alpha_{d-1}}}{(2\alpha_{d-1} + 1)!}, \qquad \gamma_0 = 1. \tag{9}$$

As a special case of Lemma 3.3 we have for the normalizing constant $c(\kappa)$, defined in (1.1), when $\alpha_0(|L|) < \infty$, that $\kappa^d c(\kappa)^{-1} = d_0(\kappa, L) = \omega_{d-1}\alpha_0(L) + o(1)$, when $\kappa \to \infty$.

Next we plug the expansion of $d_{2i}(\kappa, L)$ given in Lemma 3.3 into the expansion of the convolution given in Lemma 3.2. We will need to prove that

$$\int_0^\pi \tilde{L}_\kappa^{(s)}(\theta)\, \tilde{D}^s f(x, \theta)\, d\theta = \frac{1}{s!}\, d_s(\kappa, L)\, D^s f(x) + o(\kappa^{-s}).$$

We will arrange the terms according to the powers of κ. Define, for this purpose, for $f: S_d \to \mathbf{R}$, $s \geqslant 2$ even, and $j \geqslant 0$ integer.

$$\mathscr{D}_j^s f = \sum_{i=j}^{s/2} \frac{1}{(2i)!}\, \gamma_{s/2-i} D^{2i} f,$$

where $\gamma_{s/2-i}$ were defined in Lemma 3.3. We use the notation $\mathscr{D}_j^0 f = f$.

We will now define the smoothness class for the pointwise risk. Let $s \geqslant 2$ be even and $x_0 \in S_d$. Let $\mathbf{F}(s, x_0)$ be the set of such functions $f: S_d \to \mathbf{R}$ that $D^i f(x_0)$ is defined for $i = 1, ..., s$, for all $\xi \in T_{x_0} = \{\xi \in S_d \mid \xi \perp x_0\}$, $\partial^s / \partial \theta^s f(\phi_{x_0}^{-1}(\xi, \theta))$ is continuous as a function of $\theta \in \mathbf{R}$, $|\tilde{D}^s f(x_0, \theta)|$ is bounded for $\theta \in \mathbf{R}$, and $\lim_{\theta \to 0} \tilde{D}^s f(x_0, \theta) = D^s f(x_0)$.

LEMMA 3.4. *Let $s \geqslant 2$ be even and $f \in \mathbf{F}(s, x_0)$.*

(i) *Suppose that*

$$\int_0^\kappa t^{2i+d-1} L(t)\, dt = \alpha_{2i}(L) + o(\kappa^{2i-s}), \tag{10}$$

for $i = 1, ..., s/2$. Then

$$c(\kappa)\, f * L_\kappa(x_0) = f(x_0) + \frac{\omega_{d-1}}{d_0(\kappa, L)} \sum_{j=1}^{s/2} \kappa^{-2j} \alpha_{2j}(L)\, \mathscr{D}_1^{2j} f(x_0) + o(\kappa^{-s}).$$

(ii) *Suppose that (10) holds for $i = 0, ..., s/2$. Then*

$$\frac{\kappa^d}{\omega_{d-1}\alpha_0(L)}\, f * L_\kappa(x_0) = \alpha_0(L)^{-1} \sum_{j=0}^{s/2} \kappa^{-2j} \alpha_{2j}(L)\, \mathscr{D}_0^{2j} f(x_0) + o(\kappa^{-s}).$$

The difference between items (i) and (ii) is that in (ii) we add the assumption that $\int_0^\kappa t^{d-1} L(t)\, dt = \alpha_0(L) + o(\kappa^{-s})$. Then, by Lemma 3.3,

$$\kappa^d c(\kappa)^{-1} = d_0(\kappa, L) = \omega_{d-1} \sum_{i=0}^{s/2} \kappa^{-2i} \gamma_i \alpha_{2i}(L) + o(\kappa^{-s})$$

and (ii) follows from (i).

A class (r, s) kernel is such that in the expansion of a convolution, given in Lemma 3.4(ii), the other lower order terms will disappear exept the rth term. A class (r, s) kernel is used to estimate derivatives of order r when f is s-smooth.

DEFINITION 3.5. Let $0 \leqslant r \leqslant s$ be even. A class (r, s) kernel is a measurable function $L: [0, \infty[\to \mathbf{R}$ which satisfies

(i) $\alpha_i(|L|) < \infty$ for $i = 0, s$,

(ii) $\int_0^\kappa t^{2i+d-1} L(t)\, dt = \alpha_{2i}(L) + o(\kappa^{2i-s})$ for $i = 0, ..., s/2$,

(iii) $\alpha_r(L) \neq 0$ and $\alpha_{2i}(L) = 0$ for $i = 0, ..., r/2 - 1, r/2 + 1, ..., s/2 - 1$.

A class $(0, s)$ kernel corresponds to what is usually called a class s kernel. Note that the condition $\int_0^\kappa t^{d-1} L(t)\, dt = \alpha_0(L) + o(\kappa^{-s})$ in Definition 3.5 is

needed only in the estimation of derivatives. When L has a compact support, condition (ii) is satisfied. Thus, to construct a class s kernel, L can be fit into a polynomial model on $[0, 1]$. The next theorem gives asymptotics for the bias of the estimators $\hat{g}_n$ and $\hat{h}_{n,a}$, defined in (6) and (8).

THEOREM 3.6. *Let $0 \leqslant r < s$ be even.*

(i) *Assume that $\Delta^{r/2} f \in \mathbf{F}(s-r, x_0)$. Let L be a class $(0, s-r)$ kernel. Then*

$$E(\hat{g}_n(x_0)) = \Delta^{r/2} f(x_0) + \kappa^{r-s} \alpha_0(L)^{-1} \alpha_{s-r}(L) \mathscr{D}_1^{s-r} \Delta^{r/2} f(x_0) + o(\kappa^{r-s})$$

when $\kappa \to \infty$ and $\hat{g}_n$ was defined in (6).

(ii) *Assume that $f \in \mathbf{F}(s, x_0)$. Let L_i be a class $(2i, s)$ kernel for $i = 0, ..., r/2$. Then*

$$\mathrm{E}(\hat{h}_{n,a}(x_0)) = \sum_{i=0}^{r/2} a_i \mathscr{D}_0^{2i} f(x_0) + \kappa^{r-s} a_{r/2} \alpha_r (L_{r/2})^{-1} \alpha_s(L_{r/2}) \mathscr{D}_0^s f(x_0) + o(\kappa^{r-s})$$

when $\kappa \to \infty$ and $\hat{h}_{n,a}$ was defined in (8).

Proof. Let us first prove (i). Because the Laplace operator is symmetric,

$$\begin{aligned} \mathrm{E}(\hat{g}_n(x_0)) &= c(\kappa)\, \mathrm{E}(L^{(r,\kappa)}(x_0' X_1)) = c(\kappa) \int_{S_d} L^{(r,\kappa)}(x_0' y) f(y)\, d\mu(y) \\ &= c(\kappa) \int_{S_d} L(\kappa \arccos(x_0' y))\, \Delta^{r/2} f(y)\, d\mu(y). \end{aligned}$$

The assertion follows from Lemma 3.4(i) and from $d_0(\kappa, L) = \omega_{d-1} \alpha_0(L) + o(1)$. Let us next prove (ii). By Lemma 3.4(ii),

$$\begin{aligned} &\frac{\kappa^d}{\omega_{d-1} \alpha_{2i}(L_i)} f * (L_i)_\kappa (x_0) \\ &\quad = \kappa^{-2i} \mathscr{D}_0^{2i} f(x_0) + \kappa^{-s} \alpha_{2i}(L_i)^{-1} \alpha_s(L_i) \mathscr{D}_0^s f(x_0) + o(\kappa^{-s}), \end{aligned}$$

for $i = 0, ..., r/2$. Thus,

$$\begin{aligned} \mathrm{E}(\hat{h}_{n,a}(x_0)) &= \kappa^d\, \mathrm{E}(M_{\kappa,a}(x_0' X_1)) = \kappa^d \int_{S_d} M_{\kappa,a}(x_0' y) f(y)\, d\mu(y) \\ &= \sum_{i=0}^{r/2} a_i C_i(L_i) \kappa^{2i+d} \int_{S_d} L_i(\kappa \arccos(x_0' y)) f(y)\, d\mu(y) \\ &= \sum_{i=0}^{r/2} a_i \mathscr{D}_0^{2i} f(x_0) \\ &\quad + \kappa^{r-s} a_{r/2} \alpha_r (L_{r/2})^{-1} \alpha_s(L_{r/2}) \mathscr{D}_0^s f(x_0) + o(\kappa^{r-s}). \quad \blacksquare \end{aligned}$$

Let us define the smoothness class suitable for the study of the integrated error. Let $s \geqslant 2$ be even and $1 \leqslant p < \infty$. Let $\mathbf{F}(s, p)$ be the set of such functions $f: S_d \to \mathbf{R}$ that $\|D^i f\|_p < \infty$ for $i = 0, ..., s$, for all $x \in S_d$ and for all $\xi \in T_x = \{\xi \in S_d \mid \xi \perp x\}$, $\partial^s / \partial \theta^s f(\phi_x^{-1}(\xi, \theta))$ is continuous as a function of $\theta \in \mathbf{R}$, $\|\tilde{D}^s f(\cdot, \theta)\|_p$ is bounded for $\theta \in [0, \pi]$, and $\lim_{\theta \to 0} \|\tilde{D}^s f(\cdot, \theta) - D^s f\|_p = 0$.

Now we give a result about the L_p convergence of convolutions. The first part of the next lemma is needed when studying the variance of the kernel estimator defined in Definition 1.1. The second part assumes a smoothness condition and it is used when studying the bias of the kernel estimator.

THEOREM 3.7. *Let $f: S_d \to \mathbf{R}$ and $L: [0, \infty[\to \mathbf{R}$. Let $1 \leqslant p < \infty$.*

(i) *Assume $\|f\|_p < \infty$ and let L be a class 0 kernel. Then*

$$\lim_{\kappa \to \infty} \|c(\kappa) f * L_\kappa - f\|_p = 0.$$

(ii) *Assume $f \in \mathbf{F}(s, p)$ and let L be a class s kernel, where $s \geqslant 2$ is even. Then*

$$\lim_{\kappa \to \infty} \|\kappa^s |c(\kappa) f * L_\kappa - f| - |\alpha_0(L)^{-1} \alpha_s(L) \mathscr{D}_1^s f|\|_p = 0.$$

4. VARIANCE AND ASYMPTOTIC RISK OF THE KERNEL ESTIMATOR

Three measures of risk will be studied. First the mean squared error at a point, then the mean integrated squared error, and finally the mean integrated absolute error. Measuring loss by the L_2 error is technically easier than measuring it by the L_1 error but the L_1 error is more natural for several reasons. Firstly, it is defined for all densities. Secondly, it is invariant under scale changes. Thirdly, it is proportional to the total variation metric, that is $\int_{S_d} |f - g| \, d\mu = 2 \sup_B |\int_B f \, d\mu - \int_B g \, d\mu|$. Extensive theory of the L_1 error in the Euclidean case has been developed in Devroye and Györfi (1985), Devroye (1987).

Let us start with studying the variance of the kernel estimator. Part (i) of the following lemma concerns the estimator $\hat{g}_n$, defined in (6), and part (ii) concerns the estimator $\hat{h}_{n,a}$, defined in (8).

LEMMA 4.1. *Let $f: S_d \to \mathbf{R}$ be a bounded density which is continuous at $x_0 \in S_d$. Let $r \geqslant 0$ be even.*

(i) *Let* L *be such that* $\int_0^\infty t^{d-1}\,|L(t)|\,dt < \infty$ *and*

$$\int_0^\infty t^{m(i-r)+d-1}\,|L^{(i)}(t)|^m\,dt < \infty$$

for $m = 1, 2$, $i = 0, ..., r$. *Then*

$$Var(\hat{g}_n(x_0)) = \frac{\kappa^{2r+d}}{n}\,\frac{\beta_{r,2}(L)}{\omega_{d-1}\alpha_0(L)^2}\,f(x_0) + \frac{o(\kappa^{2r+d})}{n}$$

when $\kappa \to \infty$, *where*

$$\beta_{r,2}(L) = \int_0^\infty t^{d-1}\left\{\sum_{i=1}^{r} P_{r,i}(1)\,t^{i-r}L^{(i)}(t)\right\}^2 dt,$$

$$\beta_{0,2}(L) = \int_0^\infty t^{d-1}L^2(t)\,dt$$

where $P_{r,i}(1)$ *were defined in* (7).

(ii) *Let* $L_i\colon [0, \infty[\to \mathbf{R}$, $i = 0, ..., r/2$ *be such that* $\alpha_{2i}(|L_i|) < \infty$, $\alpha_{2i}(L_i) \neq 0$, $\alpha_0(|L_i|^m) < \infty$ *for* $m = 1, 2$. *Let* $a = (a_0, ..., a_{r/2}) \in \mathbf{R}^{r/2+1}$. *Then*

$$Var(\hat{h}_{n,a}(x_0)) = \frac{\kappa^{2r+d}}{n}\,\frac{a_{r/2}^2\alpha_0(L_{r/2}^2)}{\omega_{d-1}\alpha_r(L_{r/2})^2}\,f(x_0) + \frac{o(\kappa^{2r+d})}{n}$$

when $\kappa \to \infty$.

The optimal rate of convergence of the mean squared error of the estimator $\hat{g}_n(x_0)$ is achieved by choosing $\kappa = Cn^{1/(2s+d)}$ and then it follows from Theorem 3.6(i) and Lemma 4.1(i) that

$$\lim_{n\to\infty} n^{2(s-r)/(2s+d)}E(\hat{g}_n(x_0, \kappa) - \Delta^{r/2}f(x_0))^2$$
$$= C^{2(r-s)}\,\frac{\alpha_{s-r}(L)^2}{\alpha_0(L)^2}\,(\mathscr{D}_1^{s-r}\,\Delta^{r/2}f(x_0))^2 + C^{2r+d}\,\frac{\beta_{r,2}(L)}{\omega_{d-1}\alpha_0(L)^2}\,f(x_0).$$

This expression is minimized with respect to C, when $C = C^*$, where

$$C^* = \left[\,A(L)\,\frac{(\mathscr{D}_1^{s-r}\,\Delta^{r/2}f(x_0))^2}{f(x_0)}\right]^{1/(2s+d)}$$

and $A(L) = 2(s-r)\,\omega_{d-1}\alpha_{s-r}(L)^2/((2r+d)\,\beta_{r,2}(L))$. The expression evaluated at C^* has the value

$$(\mathscr{D}_1^{s-r}\Delta^{r/2}f(x_0))^{2(2r+d)/(2s+d)}\,f(x_0)^{2(s-r)/(2s+d)}\,\alpha_0(L)^{-2}\,\alpha_{s-r}(L)^{2(2r+d)/(2s+d)}$$
$$\times\,\beta_{r,2}(L)^{2(s-r)/(2s+d)}\,\omega_{d-1}^{2(r-s)/(2s+d)}\left(\frac{2(s-r)}{2r+d}\right)^{(4r-2s+d)/(2s+d)}.$$

The optimal rate of convergence of the mean squared error of the estimator $\hat{h}_{a,n}(x_0)$ is achieved by choosing $\kappa = Cn^{1/(2s+d)}$ and then it follows from Theorem 3.6(ii) and Lemma 4.1(ii) that

$$\lim_{n\to\infty} n^{2(s-r)/(2s+d)} E(\hat{h}_{n,a}(x_0,\kappa) - D^r f(x_0))^2$$
$$= C^{2(r-s)} \frac{[r!\,\alpha_s(L_{r/2})]^2}{\alpha_r(L_{r/2})^2} (\mathscr{D}_0^s f(x_0))^2 + C^{2r+d} \frac{\alpha_0(L_{r/2}^2)}{\omega_{d-1}\alpha_r(L_{r/2})^2} f(x_0).$$

This expression is minimized with respect to C, when $C = C^*$, where

$$C^* = \left[A(L) \frac{(\mathscr{D}_0^s f(x_0))^2}{f(x_0)} \right]^{1/(2s+d)}$$

and $A(L) = 2(s-r)\,\omega_{d-1}[r!\,\alpha_s(L_{r/2})]^2/((2r+d)\,\alpha_0(L_{r/2}^2))$. The expression evaluated at C^* has the value

$$(\mathscr{D}_0^s f(x_0))^{2(2r+d)/(2s+d)} f(x_0)^{2(s-r)/(2s+d)} \alpha_r(L_{r/2})^{-2} [r!\,\alpha_s(L_{r/2})]^{2(2r+d)/(2s+d)}$$
$$\times \alpha_0(L_{r/2}^2)^{2(s-r)/(2s+d)} \omega_{d-1}^{2(r-s)/(2s+d)} \left(\frac{2(s-r)}{2r+d}\right)^{(4r-2s+d)/(2s+d)}.$$

Part (i) of the following lemma is needed when calculating the mean integrated squared error of the kernel estimator defined in Definition 1.1 and part (ii) is needed when calculating the mean integrated absolute error.

LEMMA 4.2. *Let $f: S_d \to \mathbf{R}$ be a density function. Let $\alpha_0(|L|^i) < \infty$ for $i = 1, 2$.*

(i) *Let $\|f\|_2 < \infty$. Then,*

$$\int_{S_d} \mathrm{Var}(\hat{f}_n)\, d\mu = \frac{\kappa^d}{n} \frac{\alpha_0(L^2)}{\omega_{d-1}\alpha_0(L)^2} + \frac{o(\kappa^d)}{n},$$

when $\kappa \to \infty$.

(ii) *Secondly,*

$$\int_{S_d} \left| \sqrt{\mathrm{Var}(\hat{f}_n)} - \sqrt{\frac{\kappa^d}{n} \frac{\alpha_0(L^2)}{\omega_{d-1}\alpha_0(L)^2} f} \right| d\mu = \frac{o(\kappa^{d/2})}{n^{1/2}},$$

when $\kappa \to \infty$.

The optimal rate of convergence of the mean integrated squared error is achieved by choosing $\kappa = Cn^{1/(2s+d)}$ and then it follows from Theorem 3.7(ii) and Lemma 4.2(i) that

$$\lim_{n\to\infty} n^{2s/(2s+d)} \mathrm{E} \int_{S_d} (\hat{f}_n(\cdot, \kappa) - f)^2 \, d\mu$$

$$= C^{-2s} \frac{\alpha_s(L)^2}{\alpha_0(L)^2} \int_{S_d} (\mathscr{D}_1^s f)^2 \, d\mu + C^d \frac{\alpha_0(L^2)}{\omega_{d-1}\alpha_0(L)^2}. \tag{11}$$

This expression is minimized with respect to C, when $C = C^*$, where

$$C^* = \left[A(L) \int_{S_d} (\mathscr{D}_1^s f)^2 \, d\mu \right]^{1/(2s+d)} \tag{12}$$

and $A(L) = 2s\omega_{d-1}\alpha_s(L)^2/(d\alpha_0(L^2))$. The expression (11) evaluated at C^* has the value

$$\left[\int_{S_d} (\mathscr{D}_1^s f)^2 \right]^{d/(2s+d)} M(L)(\omega_{d-1})^{-2s/(2s+d)} (2s/d)^{(d-2s)/(2s+d)}$$

where

$$M(L) = \left[\left(\frac{\alpha_s(L)}{\alpha_0(L)} \right)^{2d} \left(\frac{\alpha_0(L^2)}{\alpha_0(L)^2} \right)^{2s} \right]^{1/(2s+d)}.$$

Minimum of $M(L)$ with respect to functions $L: [0, \infty[\to \mathbf{R}$ is achieved by

$$L_0(t) = (1 - t^s) I_{[0,1]}(t)$$

which is a class s kernel only when $s = 2$.

Next, the asymptotic of risk is calculated when the loss is the L_1 error. In the Euclidean case, the corresponding theorem has been given in Devroye and Györfi (1985, Theorem 1, Chap. 5), Hall and Wand (1988), and Holmström and Klemelä (1992). Define

$$A(t, u) = \begin{cases} u\gamma(t/u), & t \geqslant 0, \quad u > 0 \\ 0, & t \geqslant 0, \quad u = 0, \end{cases}$$

where

$$\gamma(u) = E\,|Z - u| = \sqrt{2/\pi} \left(u \int_0^u e^{-t^2/2} \, dt + e^{-u^2/2} \right),$$

$$\text{where} \quad Z \sim N(0, 1), \quad u \geqslant 0.$$

The function A is the same as in the Euclidean case. Recall that the definition of smoothness class $\mathbf{F}(s, p)$ is given before Lemma 3.7.

THEOREM 4.3. *Let $s \geqslant 2$ be even. Assume that the density $f \in \mathbf{F}(s, 1)$. Let L be a bounded class $(0, s)$ kernel. Let $\{\kappa_n\}$ be such a sequence that $\lim_{n\to\infty} \kappa_n = \infty$ and $\lim_{n\to\infty} \kappa_n^d n^{-1} = 0$. Then*

$$\mathrm{E} \int_{S_d} |\hat{f}_n - f| \, d\mu = \int_{S_d} A(\kappa_n^{-s} |z|, \kappa_n^{d/2} n^{-1/2} w) \, d\mu + o(\kappa_n^{-s}) + o(\kappa_n^{d/2} n^{-1/2}),$$

where

$$z = \alpha_0(L)^{-1} \alpha_s(L) \mathscr{D}_1^s f,$$

$$w = \left[\frac{\alpha_0(L^2)}{\omega_{d-1} \alpha_0(L)^2} f \right]^{1/2}.$$

The optimal rate of convergence is achieved by choosing $\kappa = Cn^{1/(2s+d)}$ and then it is true that

$$\lim_{n\to\infty} n^{s/(2s+d)} \mathrm{E} \int_{S_d} |\hat{f}_n - f| \, d\mu = \int_{S_d} A(C^{-s} |z|, C^{d/2} w) \, d\mu.$$

Compare this formula to formula (11). Numerical minimization of this expression with respect to C has been considered in Hall and Wand (1988), in the Euclidean case. On the other hand, because $A(t, u) \leqslant t + \sqrt{2/\pi} u$ by Devroye and Györfi (1985, p. 77), we get an upper bound

$$\int_{S_d} A(C^{-s} |z|, C^{d/2} w) \, d\mu \leqslant C^{-s} \int_{S_d} |z| \, d\mu + C^{d/2} \sqrt{2/\pi} \int_{S_d} w \, d\mu$$

which can be minimized explicitly with respect to C. For the Euclidean case, see also Holmström and Klemelä (1992, p. 257).

5. EMPIRICAL CHOICE OF THE SMOOTHING PARAMETER

In the previous section the smoothing parameter was given, which is asymptotically optimal in the mean integrated squared error sense. The question arises whether there exists an empirical choice of the smoothing parameter which is equally good in the sense that it gives the same asymptotic for the risk. In this section such data–driven device for choosing the smoothing parameter is presented.

The so-called plug-in method will be used. This method is based on plugging in estimates of the unknown quantities appearing in the formula for the asymptotically optimal smoothing parameter. This method was apparently first introduced by Woodroofe (1970), with mean squared error criterion. The plug-in method with the mean integrated squared error was apparently first introduced by Nadaraya (1974). The result given here differs from the result in the case of real line by Nadaraya (1974) in that he considered only a specific initial estimator (derivative of kernel estimator). Here sufficient conditions are formulated for any initial estimator, to be used in the plug-in method, to satisfy. Later developments have been made for example by Scott, Tapia and Thompson (1977), Park and Marron (1990), Sheather and Jones (1991), Wand and Jones (1994), Engel, Herrmann and Gasser (1995).

There are many other approaches to smoothing parameter selection, for example methods based on cross-validation and bootstrap. The plug-in method has had one of the best performances in simulation studies with Euclidean data (see Park and Marron 1990, Cao, Cuevas, and González–Manteiga 1994). With spherical data, least squares cross-validation and likelihood cross-validation methods were considered by Hall, Watson and Cabrera (1987).

It has been common to define the optimal smoothing parameter as a statistic $\hat{\kappa}$ which minimizes $\mathrm{E}(\mathrm{MISE}(\hat{\kappa}))$, where $\mathrm{MISE}(\kappa) = \mathrm{E}\int(\hat{f}_n(\cdot,\kappa)-f)^2$. However, in this study the optimal smoothing parameter is taken to be a statistic $\hat{\kappa}$ which minimizes $\mathrm{E}\int(\hat{f}_n(\cdot,\hat{\kappa})-f)^2$. A discussion of the differences between these approaches is given by Grund, Hall, and Marron (1994).

In Section 4 it was shown that if f has smoothness index s, the asymptotically optimal smoothing parameter sequence in the mean integrated squared error sense is

$$\kappa_n^* = C^* n^{1/(2s+d)}, \tag{13}$$

where C^* was defined in (12). The constant C^* depends on the unknown density through $\int_{S_d}(\mathscr{D}_1^s f)^2\,d\mu$. Let us assume that there is such estimator $\hat{\theta}_{s,n}$ for $\int_{S_d}(\mathscr{D}_1^s f)^2\,d\mu$ that,

$$\mathrm{E}(\hat{\theta}_{s,n}) - \int_{S_d}(\mathscr{D}_1^s f)^2\,d\mu = o(1) \tag{14}$$

and

$$\mathrm{E}\,|\hat{\theta}_{s,n} - E(\hat{\theta}_{s,n})|^m = o(n^{-ms/(2s+d)}), \tag{15}$$

for all $m \in \{1, 2, ...\}$. Furthermore it is assumed that

$$\hat{\theta}_{s,n} \geqslant 0 \tag{16}$$

with probability one when n is large enough. For the construction of such estimator, see Klemelä (1997). Let

$$\hat{C}_n = (A(L)(\hat{\theta}_{s,n} + b_n))^{1/(2s+d)} \wedge n^{\gamma},$$

where $b_n = o(1)$, $nb_n \to \infty$, and $\gamma > 0$ is arbitrary. The data–driven smoothing parameter is defined as

$$\hat{\kappa}_n = \hat{C}_n n^{1/(2s+d)}.$$

THEOREM 5.1. *Let $s \geqslant 2$ be even. Let $f: S_d \to \mathbf{R}$ be a continuous density, $f \in \mathbf{F}(s, 2)$, and $\int_{S_d} (\mathscr{D}^s f)^2 \, d\mu > 0$. (The definition of $\mathbf{F}$ was given before Lemma 3.7.) Assume that an estimator satisfying conditions (14), (15) and (16) exists. Let L be a class s kernel for which $\alpha_0(|L|^2) < \infty$ and $\alpha_2(|L'|) < \infty$. Put $J_1(t) = tL'(t)$ and assume that $|L| + |J_1| \leqslant J$, where $J: [0, \infty[\to \mathbf{R}$ is monotonically decreasing, bounded, and $\alpha_0(|J|^i) < \infty$ for $i = 1, 2$. Then*

$$E \int_{S_d} (\hat{f}_n(\cdot, \hat{\kappa}_n) - f)^2 \, d\mu \sim E \int_{S_d} (\hat{f}_n(\cdot, \kappa_n^*) - f)^2 \, d\mu,$$

where $a_n \sim b_n$ means $\lim_{n \to \infty} (a_n / b_n) = 1$ and κ_n^* was defined in equation (13).

Proof. Let us denote $\mu_{s,n} = \mathrm{E}(\hat{\theta}_{s,n})$. Put

$$C_n = (A(L)(\mu_{s,n} + b_n))^{1/(2s+d)} \wedge n^{\gamma} \tag{17}$$

and $\lambda_n = C_n n^{1/(2s+d)}$. From assumption (14) it follows that $\lim_{n \to \infty} C_n = C^*$. Thus, from equation (11) it is seen that

$$\mathrm{E} \int_{S_d} (\hat{f}_n(\cdot, \kappa_n^*) - f)^2 \, d\mu \sim \mathrm{E} \int_{S_d} (\hat{f}_n(\cdot, \lambda_n) - f)^2 \, d\mu$$

and it remains to prove

$$\mathrm{E} \int_{S_d} (\hat{f}_n(\cdot, \hat{\kappa}_n) - \hat{f}_n(\cdot, \lambda_n))^2 \, d\mu = o(n^{-2s/(2s+d)}).$$

Now

$$\mathrm{E}\int_{S_d}(\hat f_n(\cdot,\hat\kappa_n)-\hat f_n(\cdot,\lambda_n))^2\,d\mu$$

$$=\mathrm{E}\int_{S_d}\left[\frac{\partial}{\partial\kappa}\hat f_n(x,\kappa)\bigg|_{\kappa=\xi_n}(\hat\kappa_n-\lambda_n)\right]^2 d\mu(x)$$

$$\leqslant \mathrm{E}^{1/2}\left[\int_{S_d}\left[n^{1/(2s+d)}\frac{\partial}{\partial\kappa}\hat f_n(x,\kappa)\bigg|_{\kappa=\xi_n}\right]^2 d\mu(x)\right]^2 \mathrm{E}^{1/2}(\hat C_n-C_n)^4,$$

where ξ_n is between $\hat\kappa_n$ and λ_n with probability one. If $\mathrm{E}\,|X_n|^m=o(a_n^m)$ for $m=1,2,\ldots$, then it is denoted $X_n=o_E(a_n)$. Let us denote $a_n=n^{-s/(2s+d)}$. It can be proved that

$$\hat C_n-C_n=o_E(a_n). \tag{18}$$

It follows that

$$\mathrm{E}^{1/2}(\hat C_n-C_n)^4=o(n^{-2s/(2s+d)}).$$

It remains to prove

$$E\left\{\int_{S_d}\left[n^{1/(2s+d)}\frac{\partial}{\partial\kappa}\hat f_n(x,\kappa)\bigg|_{\kappa=\xi_n}\right]^2 d\mu(x)\right\}^2=O(1), \tag{19}$$

where ξ_n is between $\hat\kappa_n$ and λ_n with probability one. First note that

$$c'(\kappa)\sim Q(L)\,\kappa^{d-1}$$

as $\kappa\to\infty$, where $Q(L)=-\omega_{d-1}^{-1}\,\alpha_2(L')\,\alpha_0(L)^{-2}$ and $c(\kappa)\sim\kappa^d R(L)$, $R(L)=1/(\omega_{d-1}\alpha_0(L))$. It holds that

$$\frac{\partial}{\partial\kappa}\hat f_n(x,\kappa)=\frac{1}{n}\sum_{i=1}^{n}[\,c'(\kappa)\,L(\kappa\,\arccos(x'X_i))$$

$$+c(\kappa)\arccos(x'X_i)\,L'(\kappa\,\arccos(x'X_i))].$$

Let us denote $A_n=(C^*/2\leqslant\hat C_n\leqslant 2C^*)$, $s_{1n}=(C^*/2)\,n^{1/(d+4)}$ and $s_{2n}=2C^*n^{1/(d+4)}$. Let $\varepsilon>0$. In A_n, for sufficiently large n,

$$\left|\frac{\partial}{\partial\kappa}\hat{f}_n(x,\kappa)\Big|_{\kappa=\xi_n}\right|$$

$$\leqslant(|c'(\xi_n)|+|c(\xi_n)\,\xi_n^{-1}|)\,\frac{1}{n}\sum_{i=1}^{n}J(\xi_n\arccos(x'X_i))$$

$$\leqslant(|Q(L)|+|R(L)|)(1+\varepsilon)\,\frac{\xi_n^{d-1}}{n}\sum_{i=1}^{n}J(\xi_n\arccos(x'X_i))$$

$$\leqslant(|Q(L)|+|R(L)|)(1+\varepsilon)\,\frac{s_{2n}^{d-1}}{n}\sum_{i=1}^{n}J(s_{1n}\arccos(x'X_i))$$

$$\leqslant(|Q(L)|+|R(L)|)(1+\varepsilon)(C^*)^{-1}\,2^{2d-1}n^{-1/(2s+d)}$$

$$\times\frac{s_{1n}^{d}}{n}\sum_{i=1}^{n}J(s_{1n}\arccos(x'X_i)).$$

Thus, for sufficiently large n,

$$\mathrm{E}\left\{I_{A_n}\int_{S_d}\left[n^{1/(2s+d)}\,\frac{\partial}{\partial\kappa}\hat{f}_n(x,\kappa)\Big|_{\kappa=\xi_n}\right]^2 d\mu(x)\right\}^2$$

$$\leqslant((|Q(L)|+|R(L)|)(1+\varepsilon)\,2^{2d-1}(C^*)^{-1})^4$$

$$\times\,\mathrm{E}\left\{\int_{S_d}\left[\frac{s_{1n}^{d}}{n}\sum_{i=1}^{n}J(s_{1n}\arccos(x'X_i))\right]^2 d\mu(x)\right\}^2.$$

By Jensen's inequality and by Lemma 5.2, which is given after this proof,

$$\mathrm{E}\left\{\int_{S_d}\left[\frac{s_{1n}^{d}}{n}\sum_{i=1}^{n}J(s_{1n}\arccos(x'X_i))\right]^2 d\mu(x)\right\}^2$$

$$\leqslant\omega_d\int_{S_d}\mathrm{E}\left[\frac{s_{1n}^{d}}{n}\sum_{i=1}^{n}J(s_{1n}\arccos(x'X_i))\right]^4 d\mu(x)$$

$$=O((s_{1n}^{d}n^{-1})^2)+O(1)=O(1)$$

Thus,

$$\mathrm{E}\left\{I_{A_n}\int_{S_d}\left[n^{1/(2s+d)}\,\frac{\partial}{\partial\kappa}\hat{f}_n(x,\kappa)\Big|_{\kappa=\xi_n}\right]^2 d\mu(x)\right\}^2=O(1). \tag{20}$$

With probability one, for sufficiently large n,

$$\hat{\kappa}_n=n^{1/(2s+d)}\hat{C}_n\geqslant n^{1/(2s+d)}((A(L)\,b_n)^{1/(2s+d)}\wedge n^{\gamma})\to\infty$$

and thus $\xi_n \to \infty$ with probability one. Also, with probability one, for sufficiently large n, $\xi_n \leqslant n^{\gamma} n^{1/(2s+d)}$. Thus in A_n^c, for sufficiently large n, when M is a bound for $|J|$,

$$\left| \frac{\partial}{\partial \kappa} \hat{f}_n(x_0, \kappa) \Big|_{\kappa = \xi_n} \right| \leqslant M(|c'(\xi_n)| + |c(\xi_n)\, \xi_n^{-1}|)$$
$$\leqslant M(|Q(L)| + |R(L)|)(1+\varepsilon)\, \xi_n^{d-1}$$
$$\leqslant M(|Q(L)| + |R(L)|)(1+\varepsilon)\, n^{\gamma(d-1)} n^{(d-1)/(2s+d)}.$$

Thus, for sufficiently large n,

$$E \left\{ I_{A_n^c} \int_{S_d} \left[n^{1/(2s+d)} \frac{\partial}{\partial \kappa} \hat{f}_n(x, \kappa) \Big|_{\kappa = \xi_n} \right]^2 d\mu(x) \right\}^2$$
$$\leqslant (M(|Q(L)| + |R(L)|)(1+\varepsilon)\, n^{\gamma(d-1)} n^{d/(2s+d)})^4\, \omega_d^2 P(A_n^c) = o(1),$$

because, for sufficiently large n,

$$P(A_n^c) = P((\hat{C}_n < C^*/2) \cup (\hat{C}_n > 2C^*)) \leqslant P(|\hat{C}_n - C_n| > C^*/4)$$
$$\leqslant (C^*/4)^{-m}\, \mathrm{E}\, |\hat{C}_n - C_n|^m = o(n^{-ms/(2s+d)}) \tag{21}$$

for $m \in \{1, 2, ...\}$ by equation (18). Equation (19) follows from equations (20) and (21). ∎

The proof of Nadaraya (1974) differs from the proof given here in that he defines the constant C_n in equation (17) by an Euclidean equivalent of $C_n^{2s+d} = A(L)[\int_{S_d} \mu_{s,n}^2\, d\mu + b_n]$, where $\mu_{s,n} = \mathrm{E}\, \mathscr{D}^s \hat{f}_n$ and $\hat{f}_n$ is a kernel estimator. The following lemma was needed in the previous proof.

LEMMA 5.2. *Let $X_1, ..., X_n$ be i.i.d with a bounded density. Let $L: [0, \infty[\to \mathbf{R}$ be bounded and let $\alpha_0(|L|^2) < \infty$. Let $\{\kappa_n\}$ be such a sequence that $\lim_{n\to\infty} \kappa_n = \infty$ and $\lim_{n\to\infty} \kappa_n^d n^{-1} = 0$. Put $Z_{ni} = Z_{ni}(x) = \kappa_n^d L(\kappa_n \arccos(x' X_i))$, $i = 1, ..., n$. Then for $m \geqslant 4$ even,*

$$\sup_{x \in S_d} \mathrm{E} \left| \frac{1}{n} \sum_{i=1}^{n} Z_{ni}(x) - \mathrm{E} Z_{n1}(x) \right|^m = O\left(\left(\frac{\kappa_n^d}{n} \right)^{m/2} \right).$$

ACKNOWLEDGMENTS

The support and comments of Lasse Holmström are greatly acknowledged. Also the comments of a referee are acknowledged.

REFERENCES

1. Z. D. Bai, C. Radhakrishna Rao, and L. C. Zhao, Kernel estimators of density function of directional data, *J. Multivariate Anal.* **27** (1988), 24–39.
2. R. Beran, Exponential models for directional data, *Ann. Statist.* **7** (1979), 1162–1178.
3. J. Bretagnolle and C. Huber, Estimation des densités: Risque minimax, *Z. Wahrsch. Verw. Gebiete* **47** (1979), 119–137.
4. R. Cao, A. Cuevas, and W. González-Manteiga, A comparative study of several smoothing methods in density estimation, *Comp. Statist. Data Anal.* **17** (1994), 153–176.
5. L. Devroye, "A Course in Density Estimation," Birkhäuser, Boston, 1987.
6. L. Devroye and L. Györfi, "Density Estimation: The L^1 View," Wiley, New York, 1985.
7. P. Diaconis, "Group Representation in Probability and Statistics," IMS Lecture Notes, Vol. 11, 1988.
8. J. Engel, E. Herrmann, and T. Gasser, An iterative band-width selector for kernel estimation of densities and their derivatives, *J. Nonparametric Statist.* **4** (1995), 21–34.
9. N. I. Fisher, L. Lewis, and B. J. J. Embleton, "Statistical Analysis of Spherical Data," Cambridge Univ. Press, 1987.
10. B. Grund, P. Hall, and J. S. Marron, Loss and risk in smoothing parameter selection, *J. Nonparametric Statist.* **4** (1994), 107–132.
11. P. Hall, S. J. Sheather, M. C. Jones, and J. S. Marron, On optimal data-based bandwidth selection in kernel density estimation, *Biometrika* **78** (1991), 263–269.
12. P. Hall and M. Wand, Minimizing L^1 distance in nonparametric density estimation, *J. Multivariate Anal.* **26** (1988), 59–88.
13. P. Hall, G. S. Watson, and J. Cabrera, Kernel density estimation with spherical data, *Biometrika* **74** (1987), 751–62.
14. H. Hendriks, J. H. M. Janssen, and F. H. Ruymgaart, Strong uniform convergence of density estimators on compact Euclidian manifolds, *Statist. Probab. Lett.* **16** (1993), 305–311.
15. L. Holmström and J. Klemelä, Asymptotic bounds for the expected L^1 error of a multivariate kernel density estimator, *J. Multivariate Anal.* **42** (1992), 245–266.
16. J. Klemelä, "Estimation of Densities and Functionals of Densities with Spherical Data," Rolf Nevanlinna Institute Research Reports A, Vol. 16, 1997.
17. K. V. Mardia, "Statistics of Directional Data," Academic Press, London, 1972.
18. E. A. Nadaraya, On the integral mean square error of some nonparametric estimates for the density function, *Theor. Probability Appl.* **19** (1974), 133–141.
19. B. U. Park and J. S. Marron, Comparison of data-driven bandwidth selectors, *J. Amer. Statist. Assoc.* **85** (1990), 66–72.
20. Deleted in proof.
21. F. H. Ruymgaart, Strong uniform convergence of density estimators on spheres, *J. Statist. Plann. Inference* **23** (1989), 45–52.
22. D. W. Scott, R. A. Tapia, and J. R. Thompson, Kernel density estimation revisited, *Nonlinear Anal. Theory Meth. Applic.* **1** (1977), 339–372.
23. S. J. Sheather and M. C. Jones, A reliable data-based band-width selection method for kernel density estimation, *J. Roy. Statist. Soc. Ser. B* **53** (1991), 683–690.
24. M. P. Wand and M. C. Jones, Multivariate plug-in bandwidth selection, *Comp. Statist.* **9** (1994), 97–116.
25. G. S. Watson, Orientation statistics in the earth sciences, *Bull. Geol. Inst. Univ. Uppsala N.S.* **2** (1970), 73–89.

26. G. S. Watson, "Statistics on Spheres," University of Arkansas Lecture Notes in Mathematical Sciences, Vol. 6, Wiley, New York, 1983.
27. G. S. Watson and M. R. Leadbetter, On the estimation of the probability density, I, *Ann. Math. Statist.* **34** (1963), 480–491.
28. M. Woodroofe, On choosing a delta-sequence, *Ann. Math. Statist.* **41** (1970), 1665–1671.